Scientific Programming 20 (2012) 1–2
DOI 10.3233/SPR-2012-0337
IOS Press

Editorial

Biological knowledge discovery and data mining

Mohammad Al Hasan[a], Jun Huan[b], Jake Chen[c] and Mohammed J. Zaki[d]

[a] *Department of Computer and Information Science, Indiana University–Purdue University, Indianapolis, IN, USA*
[b] *Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA*
[c] *Indiana University School of Informatics, Indiana University–Purdue University, Indianapolis, IN, USA*
[d] *Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA*

Life science research has witnessed a paradigm shift over the last decade. The Human genome project, which started in nineties, completed in the year 2003, and contributed tremendously to the development of high throughput genome sequencing instruments. The cost of sequencing also dropped dramatically over the years. As a result, sequencing data from thousands of genomes, including plants, mammals and microbial genomes, are accumulating at an unprecedented rate. Besides the sequencing instruments, the technology and the cost of DNA microarrays, tandem mass spectrometers and high-power NMRs have also been improved favorably, which have made molecular biology and genetics data-rich. The ongoing influx of these data, the inherent uncertainties in data collection processes, and the gap between data collection and knowledge curation have created exciting opportunities for data mining researchers. Apparently, most of recent researches in various life-science related disciplines such as personalized genomics, functional genomics, proteomics and structural genomics are data-driven, where knowledge discovery and data mining (KDD) processes are playing increasingly important roles. While tremendous progress has been made over the years, many of the fundamental problems in bioinformatics, such as protein structure prediction, gene–environment interaction, and regulatory pathway mapping are still open. Data mining will continue to play an essential role in understanding these fundamental problems and in the development of novel therapeutic/diagnostic solutions in post-genome medicine.

Papers for this Special Issue were selected from the papers accepted for the 10th International Workshop on Data Mining in Bioinformatics (BIOKDD), which was held on August 21, 2011 in San Diego, CA, USA. To meet the acceptance criteria of the *Scientific Programming* journal, each of the papers in this Special Issue was expanded by their respective authors. Then, the selected submissions went through two rounds of reviews by at least two reviewers. We are very grateful to the anonymous reviewers in helping us select the following papers for this Special Issue.

The first paper, "Analyze influenza virus sequences using binary encoding approach", by Ham Ching Lam, Srinand Sreevatsan and Daniel Boley, presents an interesting approach for capturing mutation pattern of influenza virus sequences using principal component analysis (PCA). For this, the authors encode the virus sequences using bitvectors and obtain a pairwise alignment matrix in which each entry represents the similarity between the corresponding pair of sequences. Principal component decomposition of this matrix and subsequent plots of the top two principal components reveal interesting evolution history of various influenza viruses.

In the paper, "Mining low-variance biclusters to discover coregulation modules in sequencing datasets", Zhen Hu and Raj Bhatnagar present an algorithm for finding low-variance biclusters from dyadic data. Traditional biclustering algorithms do not seek biclusters with specifiable bounds on variance among different cell values. On the other hand, the proposed algorithm accepts an upper bound on variance and searches the enormous combinatorial space of biclusters to return those that satisfy the desired selection criteria. They also propose pruning mechanisms to reduce the search space significantly, which makes the algorithm efficient. Low-variance biclusters have applications in the domain of bioinformatics, specifically to find gene and transcription factors (TF) biclusters; their experiments show that the proposed algorithm obtains superior biclusters than its competitors when used for finding biclusters between genes and Transcription Factors.

The third paper, "Lung cancer survival prediction using ensemble data mining on SEER data", by Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi and Alok Choudhary, discusses the author's effort and experience for building a lung cancer survival prediction system using various supervised classification methods. They find that careful pre-processing of various features improves the prediction accuracy of the system significantly. They also find that, though various data points related to lung cancer are collected in clinical setting, a small number of them are sufficient to obtain an accuracy that is almost identical to what can be obtained by a full set of features. Among various classifiers, authors reported that random forest, an ensemble classification system, performs the best. As an outcome of this research, the authors also publish a publicly-accessible web-based tool that can predict lung cancer survivability over a range of five years.

To conclude, we thank the authors and the reviewers for their contribution to this Special Issue of *Scientific Programming* journal. We also thank Prof. Boleslaw Szymanski, co-EIC, for his encouragement and help for this Special Issue.